Galaxy(A Web-Based Genome Analysis Tool for Experimentalists) & Pearls(Extract From Past, Prepare For Future)

Yi Xianfu

Institute of Health Sciences

October 20, 2010

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

常见的文本操作、解决方案与弊端所在

- 文本操作
  显示内容、查看行数；
  提取数行/列；添加删除数行/列；
  排序、去冗余；
  取交集、并集、补集；
  ……

常见的文本操作、解决方案与弊端所在

- 文本操作
  显示内容、查看行数；
  提取数行/列；添加删除数行/列；
  排序、去冗余；
  取交集、并集、补集；
  ……
- 解决方案
  文本编辑器；
  Office（Excel）；
  手工处理；
  ……

常见的文本操作、解决方案与弊端所在

- 文本操作
  显示内容、查看行数；
  提取数行/列；添加删除数行/列；
  排序、去冗余；
  取交集、并集、补集；
  ……
- 解决方案
  文本编辑器；
  Office（Excel）；
  手工处理；
  ……
- 弊端所在
  大文件打开慢、打不开；
  Excel最多65536行、256列（2007版：1048576行、16384列）
  手工处理费时费力；
  ……

文本格式简介

- TSV & CSV
  TSV(Tabular): Tab Separated Values
  CSV: Comma Separated Values

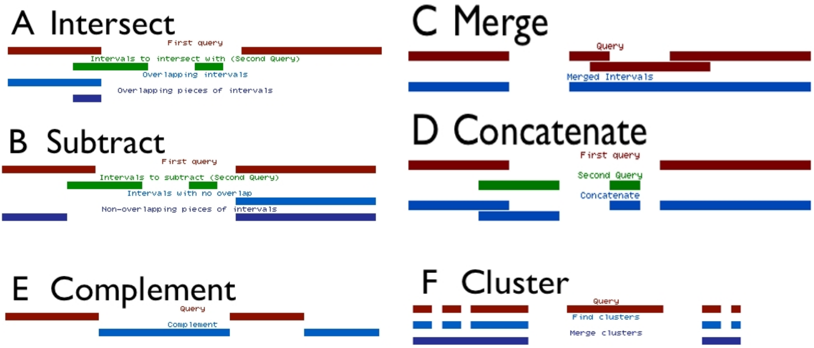
文本格式简介

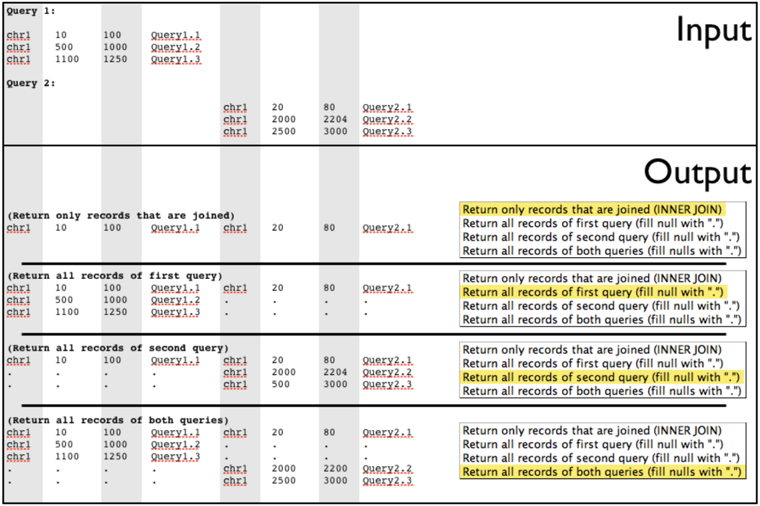
- TSV & CSV
  TSV(Tabular): Tab Separated Values
  CSV: Comma Separated Values
- FASTA, BED, GFF, FASTQ, . . .

文本格式简介

- TSV & CSV
  TSV(Tabular): Tab Separated Values
  CSV: Comma Separated Values
- FASTA, BED, GFF, FASTQ, . . .
- Interval
  Separator: Tab
  Necessary: Chromosome, Start, End
  Optional: Name, Strand, . . .

Interval Operations in Galaxy





Input

Query 1:

```
chr1   10     100    Query1.1
chr1   500    1000   Query1.2
chr1   1100   1250   Query1.3
```

Query 2:

```
chr1   20     80     Query2.1
chr1   2000   2204   Query2.2
chr1   2500   3000   Query2.3
```

Output

(Return only records that are joined)

```
chr1   10     100    Query1.1   chr1   20     80     Query2.1
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

(Return all records of first query)

```
chr1   10     100    Query1.1   chr1   20     80     Query2.1
chr1   500    1000   Query1.2   .      .      .      .
chr1   1100   1250   Query1.3   .      .      .      .
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

(Return all records of second query)

```
chr1   10     100    Query1.1   chr1   20     80     Query2.1
.      .      .      .          chr1   2000   2204   Query2.2
.      .      .      .          chr1   500    3000   Query2.3
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

(Return all records of both queries)

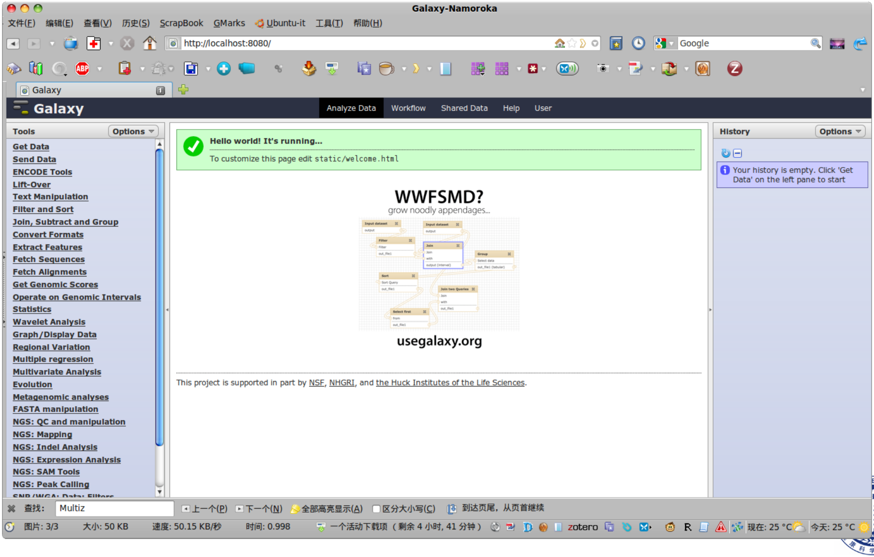
```
chr1   10     100    Query1.1   chr1   20     80     Query2.1
chr1   500    1000   Query1.2   .      .      .      .
chr1   1100   1250   Query1.3   .      .      .      .
.      .      .      .          chr1   2000   2200   Query2.2
.      .      .      .          chr1   2500   3000   Query2.3
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

Galaxy界面(localhost)



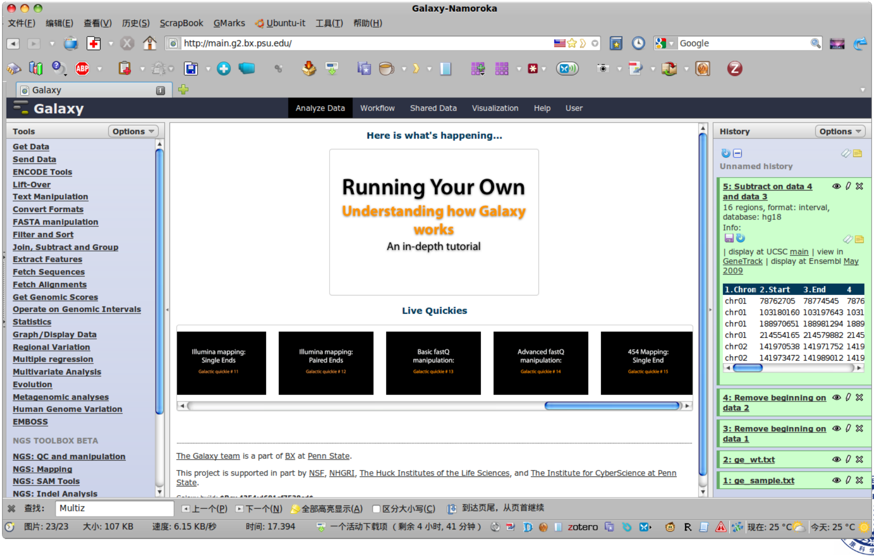
Hello world! It's running...

To customize this page edit `static/welcome.html`

This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

Your history is empty. Click 'Get Data' on the left pane to start

Galaxy界面(web)



Here is what's happening...

Live Quickies

The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

5: Subtract on data 4 and data 3

16 regions, format: interval, database: hg18

Info:

| display at UCSC main | view in GeneTrack | display at Ensembl May 2009

1.Chrom	2.Start	3.End	4
chr01	78762705	78774545	7876
chr01	103180160	103197643	1031
chr01	188970651	188981294	1889
chr01	214554165	214579882	2145
chr02	141970538	141971752	1419
chr02	141973472	141989012	1419

4: Remove beginning on data 2

3: Remove beginning on data 1

2: ge_wt.txt

1: ge_sample.txt

Four regions & Three colours

Four regions

1. Masthead: at the top

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side
3. Work area: in the middle

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side
3. Work area: in the middle
4. History panel: on the right

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side
3. Work area: in the middle
4. History panel: on the right

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side
3. Work area: in the middle
4. History panel: on the right

Three colours

- Green background: a completed query

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side
3. Work area: in the middle
4. History panel: on the right

Three colours

- Green background: a completed query
- Yellow background with rotating hourglass: a running query

Four regions & Three colours

Four regions

1. Masthead: at the top
2. Tool menu: on the left-hand side
3. Work area: in the middle
4. History panel: on the right

Three colours

- Green background: a completed query
- Yellow background with rotating hourglass: a running query
- Gray background with clock icon: a query in the queue

Several icons

- 👁 shows entire dataset in the browser window

Several icons

- 👁 shows entire dataset in the browser window
- ✏ open metadata editor

Several icons

- 👁 shows entire dataset in the browser window
- ✎ open metadata editor
- ✖ delete item from the history

Several icons

- shows entire dataset in the browser window
- open metadata editor
- delete item from the history
- save dataset to the desktop of your computer

Several icons

- 👁 shows entire dataset in the browser window
- ✎ open metadata editor
- ✖ delete item from the history
- 💾 save dataset to the desktop of your computer
- ↻ refresh OR run the job again

Several icons

- 👁 shows entire dataset in the browser window
- ✎ open metadata editor
- ✖ delete item from the history
- 💾 save dataset to the desktop of your computer
- ↻ refresh OR run the job again
- 🏷 edit history OR dataset tags

Several icons

- 👁 shows entire dataset in the browser window
- ✎ open metadata editor
- ✖ delete item from the history
- 💾 save dataset to the desktop of your computer
- 🔄 refresh OR run the job again
- 🏷 edit history OR dataset tags
- 📝 edit history OR dataset annotation

Several icons

- 👁 shows entire dataset in the browser window
- ✎ open metadata editor
- ✖ delete item from the history
- 💾 save dataset to the desktop of your computer
- 🔄 refresh OR run the job again
- 🏷 edit history OR dataset tags
- 📝 edit history OR dataset annotation
- ⊟ collapse all datasets in the history

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs

Demo

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others
3. Workflow: generate a workflow from a history in galaxy

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others
3. Workflow: generate a workflow from a history in galaxy
4. Modify: modify a parameter in the workflow in galaxy

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others
3. Workflow: generate a workflow from a history in galaxy
4. Modify: modify a parameter in the workflow in galaxy
5. Rerun: run workflows with galaxy

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others
3. Workflow: generate a workflow from a history in galaxy
4. Modify: modify a parameter in the workflow in galaxy
5. Rerun: run workflows with galaxy
6. Share: share workflows with galaxy

Demo Cont.

Share, workflow

1. Save: rename the history as “Exons and SNPs”
2. Share: share data with others
3. Workflow: generate a workflow from a history in galaxy
4. Modify: modify a parameter in the workflow in galaxy
5. Rerun: run workflows with galaxy
6. Share: share workflows with galaxy
7. Create: generate workflows from scratch with galaxy
   Find the 50 longest exons

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

参考资料

1. Home: http://galaxy.psu.edu/.
2. Wiki: http://bitbucket.org/galaxy/galaxy-central/wiki/Home.
3. Screencasts: http://main.g2.bx.psu.edu/screencast.
4. Main: http://main.g2.bx.psu.edu/.
5. Test: http://test.g2.bx.psu.edu/.

6. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome research. 2005;15(10):1451-5.
7. Taylor J, Schenck I, Blankenberg D, Nekrutenko A. Using galaxy to perform large-scale interactive data analyses. Current protocols in bioinformatics. 2007;Chapter 10:Unit 10.5.
8. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology.
9. Blankenberg D, Gordon A, Von Kuster G, et al. Manipulation of FASTQ data with Galaxy. Bioinformatics (Oxford, England). 2010;26(14):1783-5.
10. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology. 2010;Chapter 19(January):Unit 19.10.1-21.

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

三段论

最初输入

数个文件
原始格式
数据缺失

. . .

三段论

最初输入

数个文件
原始格式
数据缺失

. . .

中间处理

过滤筛选
添加删除
信息统计

. . .

三段论

最初输入

数个文件
原始格式
数据缺失

...

中间处理

过滤筛选
添加删除
信息统计

...

最终输出

数个文件
规范格式
表格图片

...

三段论

最初输入

数个文件
原始格式
数据缺失
. . .

中间处理

过滤筛选
添加删除
信息统计
. . .

最终输出

数个文件
规范格式
表格图片
. . .

Three steps all need a key file: README

1. 输入数据：数据来源、格式说明、行列解释、目的要求

三段论

最初输入

数个文件
原始格式
数据缺失
...

中间处理

过滤筛选
添加删除
信息统计
...

最终输出

数个文件
规范格式
表格图片
...

Three steps all need a key file: README

1. 输入数据：数据来源、格式说明、行列解释、目的要求
2. 处理过程：程序来源、操作步骤、参数设定、细节解释

三段论

最初输入

数个文件
原始格式
数据缺失
...

中间处理

过滤筛选
添加删除
信息统计
...

最终输出

数个文件
规范格式
表格图片
...

Three steps all need a key file: README

1. 输入数据：数据来源、格式说明、行列解释、目的要求
2. 处理过程：程序来源、操作步骤、参数设定、细节解释
3. 输出文件：格式说明、行列注解、版本控制、缩写解释

规范标准

- 文本文件格式：TSV

规范标准

- 文本文件格式：TSV
- 嵌入注释信息：位于文件顶端，以#开头

规范标准

- 文本文件格式：TSV
- 嵌入注释信息：位于文件顶端，以#开头
- 列名命名规范：若干有意义的单词，首字母大写、单词间无空格，如：GeneName

规范标准

- 文本文件格式：TSV
- 嵌入注释信息：位于文件顶端，以#开头
- 列名命名规范：若干有意义的单词，首字母大写、单词间无空格，如：GeneName
- 文本文件命名：以_分隔若干有意义的单词，并附加年月日，后缀可有可无（若有，一般为txt），如：dbSNP130_hg18_UCSC_table_20101001.txt

规范标准

- 文本文件格式：TSV
- 嵌入注释信息：位于文件顶端，以#开头
- 列名命名规范：若干有意义的单词，首字母大写、单词间无空格，如：GeneName
- 文本文件命名：以_分隔若干有意义的单词，并附加年月日，后缀可有可无（若有，一般为txt），
  如：dbSNP130_hg18_UCSC_table_20101001.txt
- 撰写README文件：清晰、全面、详细

规范标准

- 文本文件格式：TSV
- 嵌入注释信息：位于文件顶端，以#开头
- 列名命名规范：若干有意义的单词，首字母大写、单词间无空格，如：GeneName
- 文本文件命名：以_分隔若干有意义的单词，并附加年月日，后缀可有可无（若有，一般为txt），如：dbSNP130_hg18_UCSC_table_20101001.txt
- 撰写README文件：清晰、全面、详细
- 输入输出文件：避免程序无法处理的标注，可以添加一列信息进行替代

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed: [3,4]

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed: [3,4]
   - 0-based, fully-closed: [2,3]
3. Example:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed: [3,4]
   - 0-based, fully-closed: [2,3]
3. Example:
   - Genome Browser: 1-based, fully-closed:

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed: [3,4]
   - 0-based, fully-closed: [2,3]
3. Example:
   - Genome Browser: 1-based, fully-closed:
   - Table Browser: 0-based, half-open

0-based & 1-based

1. Sequence:

0-based index	0	1	2	3	4	5	6	7
sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

2. Coordinates of TT:
   - 0-based, half-open: [2,4)
     also known as: 0-based start, 1-based end
   - 1-based, fully-closed: [3,4]
   - 0-based, fully-closed: [2,3]
3. Example:
   - Genome Browser: 1-based, fully-closed:
   - Table Browser: 0-based, half-open
   - dbSNP: 0-based, half-open

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

LiftOver

- Introduction
  This tool converts genome coordinates and genome annotation files between assemblies.

LiftOver

- Introduction

  This tool converts genome coordinates and genome annotation files between assemblies.
- Websites

LiftOver

- Introduction
  This tool converts genome coordinates and genome annotation files between assemblies.
- Websites
  - LiftOver: http://hgdownload.cse.ucsc.edu/downloads.html#liftover.

LiftOver

- Introduction
  This tool converts genome coordinates and genome annotation files between assemblies.
- Websites
  - LiftOver: http://hgdownload.cse.ucsc.edu/downloads.html#liftover.
  - Web version: http://genome.ucsc.edu/cgi-bin/hgLiftOver

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

EMBOSS

- Introduction
  EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web.

EMBOSS

- Introduction
  EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web.
- Websites

EMBOSS

- Introduction
  EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web.
- Websites
  - EMBOSS: http://emboss.sourceforge.net/index.html

EMBOSS

- Introduction

  EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web.

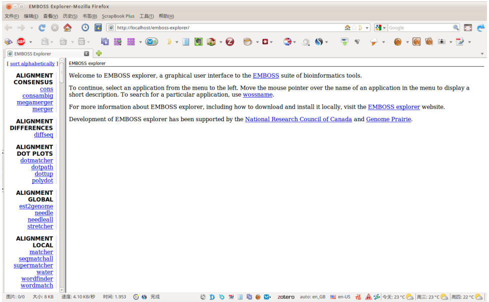
- Websites
  - EMBOSS: http://emboss.sourceforge.net/index.html
  - Jemboss: http://emboss.sourceforge.net/Jemboss/

EMBOSS

- Introduction

  EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web.
- Websites
  - EMBOSS: http://emboss.sourceforge.net/index.html
  - Jemboss: http://emboss.sourceforge.net/Jemboss/
  - Web interfaces: http://emboss.sourceforge.net/interfaces/#web

EMBOSS Explorer & Jemboss


EMBOSS Explorer-Mozilla Firefox
文件(F) 编辑(E) 查看(V) 历史(S) 书签(B) ScrapBook Plus 工具(T) 帮助(H)
http://localhost/emboss-explorer/
Google
EMBOSS Explorer
图片: 0/0 大小: 8 KB 速度: 4.10 KB/秒 时间: 1.953 完成
zotero auto: en_GB en-US 今天: 23 °C 周三: 23 °C 周四: 22 °C

[sort alphabetically]

ALIGNMENT CONSENSUS
cons
consambig
megamerger
merger

ALIGNMENT DIFFERENCES
diffseq

ALIGNMENT DOT PLOTS
dotmatcher
dotpath
dottup
polydot

ALIGNMENT GLOBAL
est2genome
needle
needleall
stretcher

ALIGNMENT LOCAL
matcher
seqmatchall
supermatcher
water
wordfinder
wordmatch

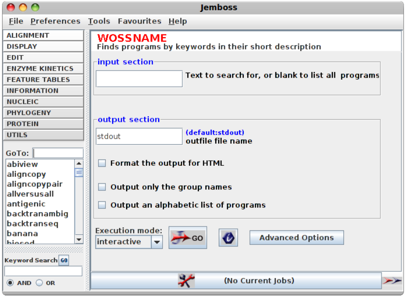
EMBOSS explorer

Welcome to EMBOSS explorer, a graphical user interface to the EMBOSS suite of bioinformatics tools.

To continue, select an application from the menu to the left. Move the mouse pointer over the name of an application in the menu to display a short description. To search for a particular application, use wossname.

For more information about EMBOSS explorer, including how to download and install it locally, visit the EMBOSS explorer website.

Development of EMBOSS explorer has been supported by the National Research Council of Canada and Genome Prairie.



WOSSNAME
Finds programs by keywords in their short description

input section
Text to search for, or blank to list all programs

output section
stdout (default:stdout)
outfile file name

□ Format the output for HTML

□ Output only the group names

□ Output an alphabetic list of programs

Execution mode: interactive

提纲

1 Galaxy

- 相关背景介绍
- Galaxy简介
- 实例演示
- 参考资料

2 Pearls

- “输入-加工-输出”三段论
- 0-based与1-based
- 基因组版本转换LiftOver
- 基因组数据处理套件EMBOSS
- FASTX-Toolkit
- Excel与TSV的转换
- 文本编辑器
- Unix与Dos的换行符转换
- 用于文本处理的shell指令

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information
  - FASTQ/A Trimmer

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information
  - FASTQ/A Trimmer
  - FASTQ/A Clipper

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information
  - FASTQ/A Trimmer
  - FASTQ/A Clipper
  - FASTQ Quality Filter

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information
  - FASTQ/A Trimmer
  - FASTQ/A Clipper
  - FASTQ Quality Filter
  - FASTQ Quality Trimmer

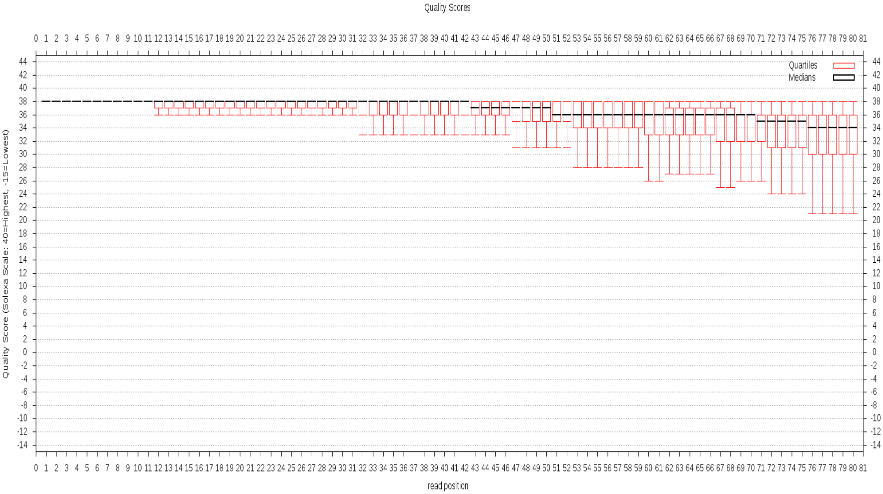
FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information
  - FASTQ/A Trimmer
  - FASTQ/A Clipper
  - FASTQ Quality Filter
  - FASTQ Quality Trimmer
  - ...

FASTX-Toolkit

- Introduction
  The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
- Available Tools
  - FASTQ-to-FASTA converter
  - FASTQ Information
  - FASTQ/A Trimmer
  - FASTQ/A Clipper
  - FASTQ Quality Filter
  - FASTQ Quality Trimmer
  - . . .
- Website
  http://hannonlab.cshl.edu/fastx_toolkit/

FASTX



Quality Scores
Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
44 42 40 38 36 34 32 30 28 26 24 22 20 18 16 14 12 10 8 6 4 2 0 -2 -4 -6 -8 -10 -12 -14

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料

2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

Excel与TSV的转换

- Excel ⇒ TSV

Excel与TSV的转换

- Excel ⇒ TSV
  - MS office, WPS: 使用另存为功能，保存类型选择文本文件(制表符分隔)。

Excel与TSV的转换

- Excel ⇒ TSV
  - MS office, WPS: 使用另存为功能，保存类型选择文本文件(制表符分隔)。
  - OpenOffice: 使用另存为功能，选择CSV文本(.csv)，同时勾选左下角的编辑筛选设置；之后，修改字段分隔符的,[逗号]为制表符，同时删除默认的文字分隔符“[双引号]即可。

Excel与TSV的转换

- Excel ⇒ TSV
  - MS office, WPS: 使用另存为功能，保存类型选择文本文件(制表符分隔)。
  - OpenOffice: 使用另存为功能，选择CSV文本(.csv)，同时勾选左下角的编辑筛选设置；之后，修改字段分隔符的,[逗号]为制表符，同时删除默认的文字分隔符“[双引号]即可。
- TSV ⇒ Excel
  选择使用Excel打开即可；之后可以另存为xls格式。

提纲

1. Galaxy
   - 相关背景介绍
   - Galaxy简介
   - 实例演示
   - 参考资料
2. Pearls
   - “输入-加工-输出”三段论
   - 0-based与1-based
   - 基因组版本转换LiftOver
   - 基因组数据处理套件EMBOSS
   - FASTX-Toolkit
   - Excel与TSV的转换
   - 文本编辑器
   - Unix与Dos的换行符转换
   - 用于文本处理的shell指令

推荐使用的编辑器

- Windows
  - Notpad++: http://notepad-plus-plus.org/
  - UltraEdit[Not FREE]: http://www.ultraedit.com/
  - SciTE: http://www.scintilla.org/SciTE.html
- Linux
  - gedit: http://projects.gnome.org/gedit/
  - Vim: http://www.vim.org/
  - Emacs: http://www.gnu.org/software/emacs/

提纲

1 Galaxy
- 相关背景介绍
- Galaxy简介
- 实例演示
- 参考资料

2 Pearls
- “输入-加工-输出”三段论
- 0-based与1-based
- 基因组版本转换LiftOver
- 基因组数据处理套件EMBOSS
- FASTX-Toolkit
- Excel与TSV的转换
- 文本编辑器
- Unix与Dos的换行符转换
- 用于文本处理的shell指令

Unix与Dos的换行符

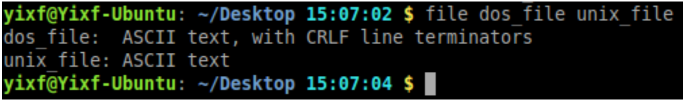
- 换行符

Unix与Dos的换行符

- 换行符
  - Windows: \r\n(回车+换行),文件尾部直接EOF(文件结束标志)

Unix与Dos的换行符

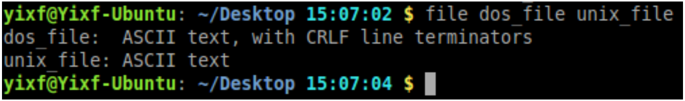
- 换行符
  - Windows: \r\n(回车+换行),文件尾部直接EOF(文件结束标志)
  - Unix: \n(仅有换行),文件最后一行也会增加该字符,然后才是EOF



```
yixf@Yixf-Ubuntu: ~/Desktop 15:07:02 $ file dos_file unix_file
dos_file:  ASCII text, with CRLF line terminators
unix_file: ASCII text
yixf@Yixf-Ubuntu: ~/Desktop 15:07:04 $ 
```

Unix与Dos的换行符

- 换行符
  - Windows: \r\n(回车+换行),文件尾部直接EOF(文件结束标志)
  - Unix: \n(仅有换行),文件最后一行也会增加该字符,然后才是EOF

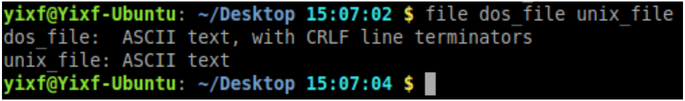


```
yixf@Yixf-Ubuntu: ~/Desktop 15:07:02 $ file dos_file unix_file
dos_file:  ASCII text, with CRLF line terminators
unix_file: ASCII text
yixf@Yixf-Ubuntu: ~/Desktop 15:07:04 $ 
```

- 转换

Unix与Dos的换行符

- 换行符
  - Windows: \r\n(回车+换行),文件尾部直接EOF(文件结束标志)
  - Unix: \n(仅有换行),文件最后一行也会增加该字符,然后才是EOF

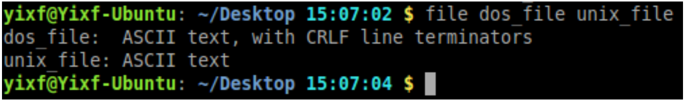


```
yixf@Yixf-Ubuntu: ~/Desktop 15:07:02 $ file dos_file unix_file
dos_file:  ASCII text, with CRLF line terminators
unix_file: ASCII text
yixf@Yixf-Ubuntu: ~/Desktop 15:07:04 $ 
```

- 转换
  - Windows: 文本编辑器

Unix与Dos的换行符

- 换行符
  - Windows: \r\n(回车+换行),文件尾部直接EOF(文件结束标志)
  - Unix: \n(仅有换行),文件最后一行也会增加该字符,然后才是EOF



```
yixf@Yixf-Ubuntu: ~/Desktop 15:07:02 $ file dos_file unix_file
dos_file:  ASCII text, with CRLF line terminators
unix_file: ASCII text
yixf@Yixf-Ubuntu: ~/Desktop 15:07:04 $ 
```

- 转换
  - Windows: 文本编辑器
  - Unix:
    Package: tofrodos
    Command: fromdos & todos

提纲

1 Galaxy
- 相关背景介绍
- Galaxy简介
- 实例演示
- 参考资料

2 Pearls
- “输入-加工-输出”三段论
- 0-based与1-based
- 基因组版本转换LiftOver
- 基因组数据处理套件EMBOSS
- FASTX-Toolkit
- Excel与TSV的转换
- 文本编辑器
- Unix与Dos的换行符转换
- 用于文本处理的shell指令

用于文本处理的shell指令

1. 显示查看
   file; cat, tac, dog[1], nl; more, less; head, tail

[1] GREEN: 需要自己下载安装

[2] http://soap.genomics.org.cn/index.html

用于文本处理的shell指令

1. 显示查看
   file; cat, tac, dog[1], nl; more, less; head, tail
2. 简单处理
   wc; sort, msort[2], uniq; split, cut, paste, colrm; diff, join, merge, comm

[1] GREEN: 需要自己下载安装

[2] http://soap.genomics.org.cn/index.html

用于文本处理的shell指令

1. 显示查看
   file; cat, tac, dog[1], nl; more, less; head, tail
2. 简单处理
   wc; sort, msort[2], uniq; split, cut, paste, colrm; diff, join, merge, comm
3. 编程处理
   tr; grep; sed, awk; perl

[1] GREEN: 需要自己下载安装

[2] http://soap.genomics.org.cn/index.html

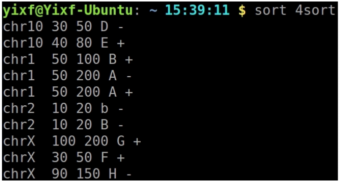
用于文本处理的shell指令

1. 显示查看
   file; cat, tac, dog[1], nl; more, less; head, tail
2. 简单处理
   wc; sort, msort[2], uniq; split, cut, paste, colrm; diff, join, merge, comm
3. 编程处理
   tr; grep; sed, awk; perl
4. 使用帮助
   man COMMAND, info COMMAND;
   COMMAND -h, COMMAND - -help

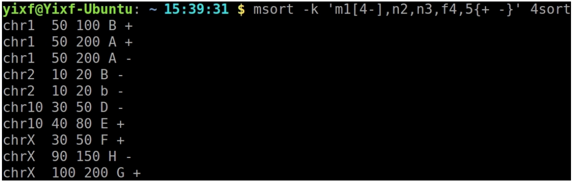
[1] GREEN: 需要自己下载安装

[2] http://soap.genomics.org.cn/index.html

sort与msort



```
yixf@Yixf-Ubuntu: ~ 15:39:11 $ sort 4sort
chr10 30 50 D -
chr10 40 80 E +
chr1  50 100 B +
chr1  50 200 A -
chr1  50 200 A +
chr2  10 20 b -
chr2  10 20 B -
chrX  100 200 G +
chrX  30 50 F +
chrX  90 150 H -
```



```
yixf@Yixf-Ubuntu: ~ 15:39:31 $ msort -k 'm1[4-],n2,n3,f4,5{+ -}' 4sort
chr1  50 100 B +
chr1  50 200 A +
chr1  50 200 A -
chr2  10 20 B -
chr2  10 20 b -
chr10 30 50 D -
chr10 40 80 E +
chrX  30 50 F +
chrX  90 150 H -
chrX  100 200 G +
```

数据资料的传递、共享、组织、控制……

- FTP/网页服务器
- 在线云存储
- 版本控制
- 数据库
- ……

Thanks for your attention!